Remarks

The Amendments

The specification and claims have been amended to italicize occurrences of the word "*Ehrlichia.*" This amendment does not narrow the claims and merely changes the font used for the word "*Ehrlichia.*"

Claims 1, 3, and 5 have been amended to delete the phrase "shown in" in favor of "consisting essentially of". This amendment is intended to provide "closed claim language" to the polypeptides themselves and NOT to the claimed compositions of matter. That is, the polypeptides consist essentially of the sequence shown in SEQ ID NO:1. The claimed compositions of matter can therefore comprise elements other than the recited polypeptides.

Claims 1, 3, and 5 have been amended to recite that the claimed variants are "amino acid substitution variants thereof that specifically bind to an anti-*Ehrlichia* antibody." Support for amino acid substitution variants can be found in the specification at, *inter alia*, page 7, line 10 through page 8, line 20. The definition of substitution variants in the specification included that the variants specifically bind to an anti-*Ehrlichia* antibody. *See e.g.,* specification, page 9, lines 8-11; page 11, lines 7-9.

Claim 5 has amended to clarify that the claim is drawn to a composition and not to a process. This is not a narrowing amendment and is made solely to clarify what is claimed. Claims 5 and 6 have also been amended to independent claims.

New claims 7 and 8 find support in the specification at, *inter alia*, Table 1.

Objections to the Specification

The specification has been objected to for the use of unitalicized term "Ehrlichia". The specification has been amended to italicize all occurrences of this word.

The specification has also been objected to as containing the use of trademarks throughout the specification. The Office Action requests correction. The M.P.E.P. states that the use of trademarks having definite meanings is permissible in patent applications as long as the proprietary nature of the marks are respected. *See* M.P.E.P. §608.01(v). The M.P.E.P. further states that trademarks should be identified by capitalizing each letter of the mark. The specification refers to the SNAP® trademark in all capital letters and also provides a registration mark where the trademark is used. As such, the proprietary nature of the mark is respected. Furthermore, the meaning of the term is definite as used in the specification. The specification teaches that:

> A preferred assay of the invention is the reversible flow chromatographic binding assay, for example a SNAP® assay. *See* U.S. Pat. No. 5,726,010. See page 13, lines 14-16.

The use of trademarks in the instant specification is proper. Applicants respectfully request withdrawal of the objections to the application.

Objection to the Claims

Claims 4-6 have been objected to for the use of unitalicized term "Ehrlichia". Claims 4-6 has been amended to italicize all occurrences of this word.

Claims 5-6 stand objected to under 37 CFR 1.75(c) as being improper dependant claims. Claims 5 and 6 have been rewritten in independent form.

Applicants respectfully request withdrawal of the objections to claims 4-6.

Information Disclosure Statement

The Office Action notes that the references listed on the information disclosure statement submitted on January 22, 2002 have been considered. However, the Office Action has requested a replacement 1449 form for this Information Disclosure Statement filed. The requested 1449 form is attached. Applicants respectfully request that the form be signed, dated and returned to the undersigned.

Rejection of Claims 1-6 Under 35 U.S.C. 112, first paragraph

Claims 1-6 stand rejected under 35 U.S.C. 112, first paragraph as allegedly lacking written description. Applicants respectfully traverse the rejection.

The Office Action asserts that the claimed variants are not adequately described by the specification.

Initially, claims 1 and 3 have been amended to clarify that the claimed variants are amino acid substitution variants of SEQ ID NO:1 that specifically bind to an anti-*Ehrlichia* antibody. The specification teaches that amino acid substitution variants of the invention can be, for example, phenotypically silent amino acid substitutions and/or conservative amino acid substitutions. The specification further provides detailed guidance on how to construct variants of SEQ ID NO:1. *See*, page 7, line 10 through page 8, line 20. *See also*, Bowie *et al.*, Science, 247:1306 (1990) (copy attached) (teaching methods of construction of variants and the tolerance of protein sequences to substitutions). The specification also teaches that polypeptides of the invention "specifically bind to an anti-*Ehrlichia* antibody". *See e.g.*, page 9, lines 8-11. The term "polypeptides of the invention" includes "variants thereof". *See e.g.*, page 11, lines 7-9.

The standard for written description requires that one of skill in the art must recognize that the applicant was in possession of the claimed genus, that is, variants of SEQ ID NO:1. Importantly:

> The written description requirement for a claimed genus may be satisfied through sufficient description of a representative number of species by actual reduction to practice, reduction to drawings, or by disclosure of relevant, identifying characteristics, i.e. structure or other physical and/or chemical properties, by functional characteristics coupled with a known or disclosed correlation between function and structure, or by a combination of such identifying characteristics, sufficient to show the applicant was in possession of the claimed genus. Guidelines for Examination of Patent Applications Under the 35 U.S.C. 112, ¶1, "Written Description" Requirement, 66 Fed. Reg. 1099, 1106 (2001) (citations omitted).

Satisfactory disclosure of a representative number of species depends on whether one of skill in the art would recognize that the applicant was in possession of the necessary common attributes or features of the elements possessed by the members of the genus in view of the species disclosed. Description of a representative number of species does not require the description to be of such specificity that it would provide individual support for each species that the genus embraces. One species can adequately support a genus.

What is a representative number of species depends on whether one of skill in the art would recognize that applicant was in possession of the necessary common attributes of features of the elements possessed by the members of the genus in view of the species disclosed or claimed. Distinguishing characteristic include:

A. partial structure;

B. physical and/or chemical properties;

C. functional characteristics;

D. known or disclosed correlation between structure and function;

E. method of making; and

F. combinations of A-E.

All of these factors, in view of the level of skill and knowledge in the art in light of and consistent with the written description, should be considered. *See* M.P.E.P. § 2163.

In the instant case, the partial structure of the claimed variants are known, *i.e.*, sequences that having at least 85% identity to SEQ ID NO:1. Therefore, the variants have about 17 amino acids in common with the 20 amino acid long SEQ ID NO:1. The physical properties and functional characteristics of the variants are known. That is, the specification teaches that the variants specifically bind to an anti-*Ehrlichia* antibody and also teaches how to test if such variants specifically bind to an anti-*Ehrlichia* antibody. *See* specification page 10, line 6 through page 11, line 6; page 11, line 21- page 16, line 8; Example 1. Methods of making the variants of SEQ ID NO:1 are well-known in the art and are described in the specification. *See e.g.* page 5, lines 7-14; page 6, line 3 through page 7, line 5; page 7, line 12 through page 9, line 7; page 18, line 19 through page 19, line 13; page 7. One of skill in the art could make and test variants of invention given the specification and the knowledge in the art.

Therefore, one of skill in the art would recognize that the Applicants were in possession of the necessary common attributes or features of the elements possessed by the members of the genus in view of the species disclosed because the partial structure, physical and/or chemical properties, functional characteristics, and methods of making the claimed variants is disclosed in the specification. The written description does not

have to be of such specificity that it would provide individual support for each species that the genus embraces.

Finally, the Office Action asserts that the species specifically disclosed are not representative of the genus because the genus is highly variant. Applicants do not agree that the genus of claimed variants is highly variant. As stated above, the claimed variants are phenotypically silent amino acid variants and conservative amino acid variants that have at least 85% identity to SEQ ID NO:1, and specifically bind to an anti-*Ehrlichia* antibody. The genus is not highly variant.

Therefore, when all factors are considered, one of skill in the art would recognize from the disclosure that the Applicants were in possession of the claimed invention. Applicants respectfully request withdrawal of the rejection.

Rejection of Claims 1-6 Under 35 U.S.C. 112, first paragraph

Claims 1-6 stand rejected under 35 U.S.C. 112, first paragraph as allegedly lacking enablement. Applicants respectfully traverse the rejection.

The Office Action asserts that the claimed variants are not enabled by the specification.

Claims 1 and 3 have been amended to clarify that the claimed variants are amino acid substitution variants of SEQ ID NO:1 that specifically bind to an anti-*Ehrlichia* antibody. The specification teaches how to make and how to use the claimed variants. The specification teaches that amino acid substitution variants of the invention can be, for example, phenotypically silent amino acid substitutions and/or conservative amino acid substitutions. The specification further provides detailed guidance on how to construct variants of SEQ ID NO:1. *See,* page 7, line 10 through page 8, line 20. *See also*, Bowie,

et al., *Science*, 247:1306 (1990) (copy attached) (teaching methods of construction of variants and the tolerance of protein sequences to substitutions). The specification also teaches that polypeptides of the invention "specifically bind to an anti-*Ehrlichia* antibody". *See e.g.*, page 9, lines 8-11. The term "polypeptides of the invention" includes "variants thereof". *See e.g.*, page 11, lines 7-9. The specification teaches how to test specific binding of a polypeptide to an anti-*Ehrlichia* antibody. *See e.g.*, Example 1. Such testing is routine to one of skill in the art.

Additionally, a structural description of the claimed variants is provided by the specification. The variants are phenotypically silent or conservative amino acid variants that have at least 85% identity to SEQ ID NO:1, and specifically bind to an anti-*Ehrlichia* antibody. Since SEQ ID NO:1 is about 20 amino acids long, an amino acid substitution variant has only about 3 amino acid substitutions at the most. One of skill in the art certainly could design, make, and test phenotypically silent and conservative amino acid variants of SEQ ID NO:1. Thus, it is trivial and routine to screen for the substitutions possible while maintaining 85% or greater identity to SEQ ID NO:1 and while maintaining binding to anti-*Ehrlichia* antibodies, according to the specification. The test of enablement is not merely quantitative, since a considerable amount of experimentation is permissible, if it is merely routine, or if the specification in question provides a reasonable amount of guidance with respect to the direction in which the experimentation should proceed. " *In re Wands*, 8 USPQ2d 1400, 1404 (Fed. Cir. 1988) (citing *In re Angstadt*, 190 USPQ 214, 217-19 (CCPA 1976)); M.P.E.P. §2164.06. Time and expense are merely factors in this consideration and are not the controlling factors. *United States v. Telectronics Inc.*, 8 USPQ2d 1217, 1223 (Fed. Cir. 1988). Therefore, one of skill in

the art could make the claimed variants using only routine experimentation. Once the claimed variants are made, one of skill in the art can use them to, *inter alia*, detect the presence of anti-*Ehrlichia* antibodies.

The Office Action further asserts that it is not routine to screen multiple substitutions with a reasonable expectation of success in obtaining similar anti-*Ehrlichia* antibody binding. The Office Action asserts that the expectation in obtaining similar anti-*Ehrlichia* antibody binding is limited in any polypeptide and the result of such modifications is unpredictable. The Office Action asserts that one of skill in the art would not expect any tolerance to multiple substitutions. A reasonable expectation of success is not the standard for enablement.

Applicants remind the Office that the standard for enablement is whether one reasonably skilled in the art (1) could make and use the invention (2) from the disclosures in the patent coupled with information known in the art (3) without undue experimentation. As taught in the specification and described above, one of skill in the art could make the claimed variants while maintaining binding to an anti-*Ehrlichia* antibody, without undue experimentation. *See e.g.*, specification at page 5, line 6 through page 11, line 20. One of skill in the art could make and use the claimed variants in light of the specification and knowledge in the art, without undue experimentation. Therefore, the claims are enabled. Applicants respectfully request withdrawal of the rejection.

Applicants respectfully request withdrawal of the rejection.

Rejection of Claims 3 and 5 Under 35 U.S.C. 112, second paragraph

Claims 3 and 5 stand rejected under 35 U.S.C. 112, second paragraph as allegedly indefinite. Applicants respectfully traverse the rejection.

The Office Action asserts that claim 3 is indefinite because it recites "a polypeptide shown in SEQ ID NO:1." This language has been amended.

The Office Action asserts that claim 5 is indefinite because it is not clear if a product or process is claimed. Claim 5 has been amended to clarify that a product is being claimed.

The Office Action asserts that claim 5 is indefinite for the use of the term "under conditions." The claim recites that certain polypeptides that specifically bind to an anti-*Ehrlichia* antibody, are contacted with a test sample suspected of comprising antibodies to *Ehrlichia*, under conditions that allow polypeptide/antibody complexes to form. The claim is describing an extremely well known method of detecting the presence of antibodies to a bacterial pathogen comprising the detection of polypeptide/antibody complexes. One of skill in the art, given the specification, which includes working examples of such detection, would clearly understand the meaning of "under conditions" that allow polypeptide/antibody complexes to form because one of skill in the art would be very familiar with such methods. The claim is therefore definite and Applicants respectfully request withdrawal of the rejection.

Rejection of Claims 1-3 Under 35 U.S.C. 102(b)

Claims 1-3 stand rejected under 35 U.S.C. 102(b) as allegedly anticipated by Rikihisa et al., WO 99/13720. Applicant respectfully traverse the rejection.

The Office Action asserts that Rikihisa teaches the polypeptide shown in SEQ ID NO:1 at Figure 21B. However, Rikihisa teaches a whole *E. canis* P30-1 protein ("[t]he p30-1 polynucleotide encodes a P30-1 protein of *E. canis* having a molecular weight of about 28.0 kDa and an amino acid sequence which is at least 85% homologous to the

amino acid sequence shown in FIG. 21B." See page 7, first full paragraph). Figure 21B shows the amino acid sequence of a whole *E. canis* P30-1 protein.

Under 35 U.S.C. § 102, a claim is anticipated only if each and every element as set forth in the claim is found in a single art reference. *Verdegaal Bros. v. Union Oil Co.*, 2 USPQ2d 1051, 10533 (Fed. Cir. 1987); *In re Recombinant DNA Technology Patent and Contract Litigation*, 30 USPQ2d 1881, 1885 (S.D. Ind.1993) ("A patent is anticipated only if all the elements and limitations of the claims are found within a single, prior art reference."); *Structural Rubber Products Co. v. Park Rubber Co.*, 223 USPQ 1264, 1270 (Fed. Cir. 1984) (All elements of the claimed invention must be contained in a single prior art disclosure and must be arranged in the prior art disclosure as in the claimed invention); M.P.E.P § 2131. Furthermore, no difference may exist between the claimed invention and the reference disclosure, as viewed by a person of ordinary skill in the field of invention. *In re Recombinant DNA Technology Patent and Contract Litigation*, 30 USPQ2d 1881, 1885 (S.D. Ind.1993). Also, the identical invention must be described or shown in as complete detail as is contained in the claim. *Richardson v. Suzuki Motor Co.*, 9 USPQ2d 1913, 1920 (Fed. Cir. 1989); *Chester v. Miller*, 15 USPQ2d 1333 (Fed. Cir. 1990); M.P.E.P. § 2131.

Rikihisa does not teach or suggest an element of the claims, that is, a polypeptide consisting essentially of SEQ ID NO:1 and amino acid substitution variants thereof. Therefore, Rikihisa cannot anticipate the claims. The Office Action appears to assert, however, that a teaching of polypeptides consisting essentially of SEQ ID NO:1 and substitution variants are inherently present in Rikihisa.

The fact that a certain characteristic may occur or be present in a prior art reference is not sufficient to establish the inherency of that characteristic. *In re Rijckaert*, 9 F.3d 1531, 1534, 28 USPQ2d 1955, 1957 (Fed. Cir. 1993); *In re Oelrich*, 666 F.2d 578, 581-82, 212 USPQ 323, 326 (CCPA 1981). "To establish inherency, the extrinsic evidence 'must make clear that the missing descriptive matter is necessarily present in the thing described in the reference, and that it would be so recognized by persons of ordinary skill. Inherency, however, may not be established by probabilities or possibilities. The mere fact that a certain thing may result from a given set of circumstances is not sufficient.' " *In re Robertson*, 49 USPQ2d 1949, 1950-51 (Fed. Cir. 1999) (citations omitted); MPEP §2112.01. "In relying upon the theory of inherency, the examiner must provide a basis in fact and/or technical reasoning to reasonably support the determination that the allegedly inherent characteristic necessarily flows from the teachings of the applied prior art." *Ex parte Levy,* 17 USPQ2d 1461, 1464 (Bd. Pat. App. & Inter. 1990) (emphasis in original); MPEP §2112.01.

The Office has not provided a basis in fact and/or technical reasoning to show that the allegedly inherent characteristic necessarily flows from the teachings of the applied prior art. Rikihisa does not teach or suggest the use of polypeptide fragments in devices and in particular does not teach or suggest the particular fragment shown in SEQ ID NO:1. Nor has that Office Action alleged that the whole recombinant protein antigens in Rikihisa would be fragmented in any way.

Additionally, the claimed compositions of matter provide greater sensitivity than the reagents taught in Rikihisa (*i.e.,* whole, recombinant proteins). See attached declaration of Dr. Chandrashekar, paragraphs 2-3 and 6-7. Therefore, the claimed

compositions of matter differ from those of Rikihisa because they provide greater sensitivity than those described in Rikihisa.

Rikihisa does not anticipate claims 1-3 because Rikihisa does not teach, suggest, or inherently disclose each and every element of claims 1-3. Applicants respectfully request withdrawal of the rejection.

Rejection of Claims 1-6 Under 35 U.S.C. 103(a)

Claims 1-6 stand rejected under 35 U.S.C. 103(a) as allegedly obvious over Rikihisa *et al.,* WO 99/13720, in view of Waner *et al.* Applicants respectfully traverse the rejection.

The Office Action asserts that Rikihisa teaches diagnostic tools for sero-diagnosing ehrlichiosis in mammals and the isolated polypeptide shown in SEQ ID NO:1. The Office Action further asserts that Waner teaches a label that indicates the use of the composition of matter or the article of manufacture.

However, as discussed above, Rikihisa does not teach or suggest isolated polypeptides consisting essentially of SEQ ID NO:1. Waner does not correct the defects of the primary reference by teaching the elements missing from Rikihisa. Since the combination of references does not teach or suggest every element of the claims, they cannot render the claims obvious. Applicants respectfully request withdrawal of the rejection.

Respectfully submitted,

Dated: 2/17/04

Lisa M.W. Hillman
Registration No. 43,673

McDonnell, Boehnen,
Hulbert & Berghoff
300 S. Wacker Drive
Chicago, IL 60606
(312) 913-0001

Deciphering the message in protein sequences: tolerance to amino acid substitutions.

by James U. Bowie, John F. Reidhaar-Olson, Wendell A. Lim and Robert T. Sauer

Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

THE GENOME IS MANIFEST LARGELY IN THE SET OF PROTEINS that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to function and carry out the instructions of the genome. Thus, comprehending the rules that relate amino acid sequence to structure is fundamental to an understanding of biological processes. Because an amino acid sequence contains all of the information necessary to determine the structure of a protein [1], it should be possible to predict structure from sequence, and subsequently to infer detailed aspects of function from the structure. However, both problems are extremely complex, and it seems unlikely that either will be solved in an exact manner in the near future. It may be possible to obtain approximate solutions by using experimental data to simplify the problem. In this article, we describe how an analysis of allowed amino acid substitutions in proteins can be used to reduce the complexity of sequences and reveal important aspects of structure and functions.

Methods for Studying Tolerance to

Sequence Variation

There are two main approaches to studying the tolerance of an amino acid sequence to change. The first method relies on the process of evolution, in which mutations are either accepted or rejected by natural selection. This method has been extremely powerful for proteins such as the globins or cytochromes, for which sequences from many different species are known [2-7]. The second approach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible [3, 8-11]. The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions [2-4, 11]. For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in lac repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent [11]. At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for lac repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a 10.sup.4.-fold reduction in activity [12]. A similar loss of activity occurs in [lambda] repressor when a DNA binding residue is changed from Asn to Asp [13]. To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity [10, 14-16]. Hence, many of the residues that are conserved in sets of related sequences play structural roles.

Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved [6]. Similar results have been seen for a number of protein families [2, 4, 5, 7, 17, 18). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in [lambda] repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the protein [9]. These substitutions were identified by a functional selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried

Deciphering the message in protein sequences: tolerance to amino acid substitutions.

by the residues that are solvent inaccessible.

Constraints on Core Sequences

Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability [19]. For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the NH.sub.2.-terminal domain of [lambda] repressor [20]. The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited [21]. Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied [22].

The cores of most proteins are quite closely packed [23], but some volume changes are acceptable. In [lambda] repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core position in the appropriate sequence contexts. Large volume changes at individual buried sites have also been observed in phylogenetic studies, where it has been noted that the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion [5, 7, 17]. Rather, local volume changes are accommodated by conformational changes in nearby side chains and by a variety of backbone movements.

The Informational Importance of the Core

With occasional exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the core is composed of side chains that can assume only a limited number of conformations [24], efficient packing must be maintained without steric clashes. How important are hydrophobicity, volume, and steric complementarity in determining whether a given sequence can form an acceptable core? Each factor is essential in a physical sense, as a stable core is probably unable to tolerate unsatisfied hydrogen bonding groups, large holes, or steric overlaps [25]. However, in an informational sense, these factors are not equivalent. For example, in experiments in which three core residues of [lambda] repressor were mutated simultaneously, volume was a relatively unimportant informational constraint because three-quarters of all possible combinations of the 20 naturally occurring amino acids had volumes within the range tolerated in the core, and yet most of these sequences were unacceptable [20]. In contrast, of the sequences that contained only the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford et al., in an extensive analysis of globin sequences [4], found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from

Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of

Deciphering the message in protein sequences: tolerance to amino acid substitutions.

different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in [lambda] repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of active sequences, but it is not possible to decide whether a side chain is functionally or structurally important just because it is invariant or conserved. To make this distinction requires an independent assay of protein folding. The ability of a mutuant protein to maintain a stably folded structure can often be measured by biophysical techniques, by susceptibility to intracellular proteolysis [26], or by binding to antibodies specific for the native structure [27, 28]. In the latter cases, it is possible to screen proteins in mutated clones for the ability to fold even if these proteins are inactive. Sets of sequences that allow formation of a stable structure can then be compared to the sets that allow both folding and function, with the active site or binding residues being those that are variable in the set of stable proteins but invariant in the set of functional proteins. The DNA-binding residues of Arc repressor were identified by this method [8]. The receptor-binding residues of human growth hormone were also identified by comparing the stabilities and activites of a set of mutant sequences [28]. However, in this case, the mutants were generated as hybrid sequences between growth hormone and related hormones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known [29, 30]. However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have been used to combine such information into more appropriately weighted sequence searches and alignments [31]. These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 [29]. Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure [32].

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected [33, 34]. These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences [6]. This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots for the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain main proteins (Fig. 4B) [35]. The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure [34]. The amphipathic character of the three [alpha]-helical regions in the Antennapedia protein [36] is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of

Deciphering the message in protein sequences: tolerance to amino acid substitutions.

Arc repressor, a small DNA-binding binding protein, was recently predicted by a similar method [8] and confirmed by nuclear magnetic resonance studies [37].

The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of difficult structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion for their structural relatedness [38]. In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure [4, 29, 38, 39]. Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure de novo. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple sequences composed only of H (hydrophobic) and P (polar) groups on two-dimensional lattices [40]. An example of such a representation is shown in Fig. 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice, and two residues cannot occupy the same space. Free energies of particular conformations are evaluated with a single term, an attraction of H groups. By considering chains of ten residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence or conformation space could be explored. The significant results were as follows: (i) not all sequences can fold into a "native" structure and only a few sequences form a unique native structure; (ii) the probability that a sequence will adopt a unique native structure increases with chain length; and (iii) the native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. Although the gap between these two-dimensional simulations and three-dimensional structures is large, the use of simple rules and sequence representations yields results similar to those expected for real proteins. Three-dimensional lattice methods are also beginning to be developed and evaluated [41].

Summary

There is more information in a set of related sequences than in a single sequence. A number of practical applications arise from an analysis of the tolerance of residue positions to change. First, such information permits the evaluation of a residue's importance to the function and stability of a protein. This ability to identify the essential elements of a protein sequence may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

REFERENCES AND NOTES

[1] C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, Cold Spring Harbor Symp. Quant. Biol. 28, 439 (1963); C. B. Anfinsen, Science 181, 223 (1973).

[2] R. E. Dickerson, Sci. Am. 242, 136 (March 1980).

[3] M. D. Hampsey, G. Das, F. Sherman, FEBS Lett. 231, 275 (1988).

[4] D. Bashford, C. Chothia, A. M. Lesk, J. Mol. Biol. 196, 199 (1987).

[5] A. M. Lesk and C. Chothia, ibid. 136, 225 (1980).

[6] M. F. Perutz, J. C. Kendrew, H. C. Watson, ibid. 13,

Deciphering the message in protein sequences: tolerance to amino acid substitutions.

669 (1965).

[7] C. Chothia and A. M. Lesk, Cold Spring Harbor Symp. Quant. Biol. 52, 399 (1965).

[8] J. U. Bowie and R. T. Sauer, Proc. Natl. Acad. Sci. U.S.A. 86, 2152 (1989).

[9] J. F. Reidhaar-Olson and R. T. Sauer, Science 241, 53 (1988); Proteins Struct. Funct. Genet., in press.

[10] D. Shortle, J. Biol. Chem. 264, 5315 (1989).

[11] J. H. Miller et al., J. Mol. Biol. 131, 191 (1979).

[12] S. Sprang et al., Science 237, 905 (1987); C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, ibid., p. 909.

[13] H. C. M. Nelson and R. T. Sauer, J. Mol. Biol. 192, 27 (1986).

[14] M. H. Hecht, J. M. Sturtevant, R. T. Sauer, Proc. Natl. Acad. Sci. U.S.A. 81, 5685 (1984).

[15] T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, Biochemistry 26, 3754 (1987).

[16] D. Shortle and A. K. Meeker, Proteins Struct. Funct. Genet. 1, 81 (1986).

[17] A. M. Lesk and C. Chothia, J. Mol. Biol. 160, 325 (1982).

[18] W. R. Taylor, ibid. 188, 233 (1986).

[19] W. Kauzmann, Adv. Protein Chem. 14, 1 (1959); R. L. Baldwin, Proc. Natl. Acad. Sci. U.S.A. 83, 8069 (1986).

[20] W. A. Lim and R. T. Sauer, Nature 339, 31 (1989); in preparation.

[21] Lesk and Chothia (5) have argued that a protein core composed solely of hydrogen-bonded residues would also be inviable on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.

[22] T. M. Gray and B. W. Matthews, J. Mol. Biol. 175, 75 (1984); E. N. Baker and R. E. Hubbard, Prog. Biophys. Mol. Biol. 44, 97 (1984).

[23] F. M. Richards, J. Mol. Biol. 82, 1 (1974).

[24] J. W. Ponder and F. M. Richards, ibid. 193, 775 (1987).

[25] J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, Biochemistry 28, 4914 (1989); W. S. Sandberg and T. C. Terwilliger, Science 245, 54 (1989).

[26] A. A. Pakula and R. T. Sauer, Proteins Struct. Funct. Genet. 5, 202 (1989).

[27] B. C. Cunningham and J. A. Wells, Science 244, 1081 (1989); R. M. Breyer and R. T. Sauer, J. Biol. Chem. 264, 13348 (1989).

[28] B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, Science 243, 1330 (1989).

[29] L. H. Pearl and W. R. Taylor, Nature 329, 351 (1987).

[30] W. J. Brown et al., J. Mol. Biol. 42, 65 (1969); J. Greer, ibid. 153, 1027 (1981); J. M. Berg, Proc. Natl. Acad. Sci. U.S.A. 85, 99 (1988).

[31] W. R. Taylor, Protein Eng. 2, 77 (1988).

[32] M. A. Navia et al., Nature 337, 615 (1989).

[33] M. Schiffer and A. B. Edmundson, Biophys. J. 7, 121 (1967); V. I. Lim, J. Mol. Biol. 88, 857 (1974); ibid., p. 873.

[34] D. Eisenberg, R. M. Weiss, T. C. Terwilliger, Nature 299, 371 (1982); D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, J. Mol. Biol. 179, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, Proc. Natl. Acad. Sci. U.S.A. 81, 140 (1984).

[35] T. R. Burglin, Cell 53, 339 (1988).

[36] G. Otting et al., EMBO J. 7, 4305 (1988).

[37] J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, Biochemistry 28, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. Patel, ibid., p. 9813.

[38] R. M. Sweet and D. Eisenberg, J. Mol. Biol. 171, 479 (1983).

[39] J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, Proteins Struct. Funct. Genet., in preparation.

[40] K. F. Lau and K. A. Dill, Macromolecules 22, 3986 (1989).

[41] A. Sikorski and J. Skolnick, Proc. Natl. Acad. Sci.

Deciphering the message in protein sequences: tolerance to amino acid substitutions.

U.S.A. 86, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, Biopolymers 26, 937 (1987); D. G. Covell and R. L. Jernigan, Biochemistry, in press.

[42] B. Lee and F. M. Richards, J. Mol. Biol. 55, 379 (1971).

[43] S. R. Jordan and C. O. Pabo, Science 242, 893 (1988).

[44] R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).

[45] J.-L. Fauchere and V. Pliska, Eur. J. Med. Chem.-Chim. Ther. 18, 369 (1983).

[46] We thank C. O. Pabo and S. Jordan for coordinates of the [NH.sub.2]-terminal domain of [lambda] repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).

(*) Present address: Department of chemistry and Biochemistry and the Molecular Biology Institute, University of California, Los Angeles, Los Angeles, CA 90024.

The authors are in the Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139.